Lecture 9: Heteroskedasticity & Autocorrelation

Prepared for ECON 5033

Eric S. Lin

Department of Economics, National Tsing Hua University

December 10, 2024

Basics

- Heteroskedasticity means that the diagonal elements in variance covariance matrix are variant across observations.
- $\mathrm{var}[\varepsilon_i] = \sigma^2$ vs. $\mathrm{var}[\varepsilon_i] = \sigma_i^2$
- Covariance matrix: $\mathrm{E}[\varepsilon\varepsilon'] = \sigma^2\Omega = \Sigma = \mathrm{diag}[\sigma_i^2]$
- Consequences of heteroskedasticity:
    - coefficient estimators
    - standard errors
    - testing

Grouped Data

- It is common in the grouped data. To see this, consider a regression using averages with different numbers of observations (say individual i in county c),

$$Y_{ic} = X_{ic}'\beta + \varepsilon_{ic}.$$

- We have data ONLY at the group level,

$$\bar{Y}_s = \frac{\sum_{i=1}^{N_c} Y_{ic}}{N_c},\ \bar{X}_c = \frac{\sum_{i=1}^{N_c} X_{ic}}{N_c},$$

and regress $\bar{Y}_c$ on $\bar{X}_c$ for $c = 1, ..., C$,

$$\bar{Y}_c = \bar{X}_c'\beta + \bar{\varepsilon}_c,$$

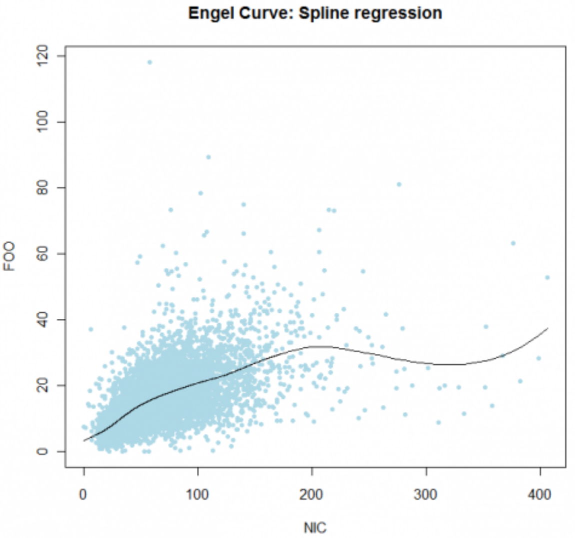
in which case,

$$\text{var}\left[\bar{Y}_c\right] = \frac{\sigma^2}{N_c} = \sigma_c^2.$$

Engel Curve

- Another leading example is the upward sloping Engel curve.
- Consider the case of regressing food expenditure (Y_i) on income (X_i).
- On average higher income corresponds with higher expenditure on food.
- In addition, one could expect that the variation of food expenditure among high income households is much larger than the variation among lower income households.
- If you do the variance plot, the cone-shaped graph suggests the possible heteroskdasticity.
- One will need a formal test for heteroskedasticity.
- But now let's suppose that heteroskedasticity is an issue!

Engel Curve


Engel Curve: Spline regression
FOO
NIC
0
20
40
60
80
100
120
0
100
200
300
400

More Examples

Example

Consider a supply function:

$$Q_i = \beta_1 + \beta_p P_i + \beta_s S_i + \varepsilon_i,$$

where P_i is price and S_i is some measure of size of the ith firm. One might suppose that unobservable factors (e.g., talent of managers, degree of coordination between production units, etc.) account for the error term ε_i. If there is more variability in these factors for large firms than for small firms, then ε_i may have a higher variance when S_i is high than when it is low.

More Examples

Example

Consider an individual demand function:

$$Q_i = \beta_1 + \beta_p P_i + \beta_{inc} Income_i + \varepsilon_i,$$

where P_i is price and $Income_i$ is income for the ith individual. In this case, ε_i can reflect variations in preferences. There are more possibilities for expression of preferences when one is rich, so it is possible that the variance of ε_i could be higher when $Income_i$ is high.

More on Causes of Heteroskedasticity

- Suppose that the slope coefficient varies across i:

$$Y_i = \alpha + X_i\beta_i + \varepsilon_i$$

- Suppose that β_i varies randomly around the fixed parameter β :

$$\beta_i = \beta + \eta_i$$

- So we have:

$$\begin{aligned} Y_i &= \alpha + (\beta + \eta_i)X_i + \varepsilon_i \\ &= \alpha + \beta X_i + v_i \\ v_i &= \eta_i X_i + \varepsilon_i \end{aligned}$$

- It is clear to see that the error term v_i varies with X_i, $\text{var}[v_i] = \sigma^2(X_i)$

Causes of Heteroskedasticity

- Suppose that the true model is given by:

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i$$

- However, if we ignore the variable Z_i and perform the following:

$$\begin{aligned} Y_i &= \alpha + \beta X_i + v_i \\ &= \alpha + \beta X_i + (\gamma Z_i + \varepsilon_i), \end{aligned}$$

  where $v_i = \gamma Z_i + \varepsilon_i$
- It is clear to see that the error term v_i varies with Z_i, $\text{var}[v_i] = \sigma^2(Z_i)$
- Furthermore, if this is the case, $\hat{\beta}_{OLS}$ will be an inconsistent estimator
- "Omitted Variable Bias (OVB)"

Causes of Heteroskedasticity

- Suppose that the true model is given by:

$$Y_i = \alpha + \beta X_i^2 + \varepsilon_i$$

- However, if we ignore non-linearity in $X_i$and perform the following:

$$\begin{aligned} Y_i &= \alpha + \beta X_i + v_i \\ &= \alpha + \beta X_i + [\beta(X_i^2 - X_i) + \varepsilon_i], \end{aligned}$$

  where $v_i = \beta(X_i^2 - X_i) + \varepsilon_i$

- It is clear to see that the error term v_i varies with X_i, $\text{var}[v_i] = \sigma^2(X_i)$
- The residuals in the regression will capture such non-linearity and its variance will be affected accordingly

Form of Heteroskedasticity

- Recall the heteroskedasticity is given by:

$$\mathrm{var}[\varepsilon] = \sigma^2\Omega$$

- Also note that one can do the following:

$$\Omega^{-1} = \omega'\omega$$

- The Ω can be written as

$$\begin{bmatrix} \sigma_1^2 & \ldots & 0 & 0 \\ \ldots & \sigma_2^2 & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_n^2 \end{bmatrix} = \mathrm{diag}[\sigma_i^2]$$

Weighted Transformation

- If we know the structure of heteroskedasticity, one could perform the transformation procedure to obtain GLS estimator.
- The transformation matrix ω is $\text{diag}(\sigma_i^{-1})$, with $\sigma^2 = 1$.
- The transformed model is,

$$\frac{Y_i}{\sigma_i} = \left(\frac{X_i}{\sigma_i}\right)' \beta + \frac{\varepsilon_i}{\sigma_i}$$
$$Y_i^* = X_i^{*\prime} \beta + \varepsilon_i^*.$$

- We weight the data (Y_i, X_i') by σ_i, the GLS estimator is also called Weighted Least Square estimator (WLS).
- WLS is given by,

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} X_i Y_i\right).$$

How to Test for Heteroskedasticity

- Several tests for detecting heteroskedasticity include:
  - Breusch-Pagan LM test
  - Koenker's test
  - White's general test
  - and many others
- First three tests are popular in econometrics

Breusch-Pagan LM Test

- Proposed by Breusch and Pagan (1979)
- Assume heteroskedasticity to be a function of the some exogenous variables that could include functions of the regressors,

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i^2 = h(Z_i'\alpha).$$

- h is any non-negative function and assume that,

$$Z_i'\alpha = \alpha_o + \alpha_1 Z_{1i} + ... + \alpha_p Z_{pi}.$$

- The BP test does not require us to specify the unknown, continuously differentiable function $h(\cdot)$.
- The heteroskedasticity test is then,

$$H_o : \alpha_1 = \alpha_2 = ... = \alpha_p = 0$$

Breusch-Pagan LM Test

- The log-likelihood is of the form,

$$\ln \mathcal{L}(\beta, \alpha) = -\frac{1}{2}\sum_{i=1}^{n} \ln\left[h\left(Z_i'\alpha\right)\right] - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left(Y_i - X_i'\beta\right)^2}{h\left(Z_i'\alpha\right)}.$$

- To construct the LM test, let's look at the (unrestricted) FOCs evaluated at restricted estimators.

Breusch-Pagan LM Test

- FOCs:

$$\frac{\partial \ln \mathcal{L}(\hat{\beta}, \hat{\alpha}_o, 0)}{\partial \beta} = \sum_{i=1}^{n} \frac{X_i(Y_i - X_i'\hat{\beta})}{h(\hat{\alpha}_o)} = 0$$

$$\frac{\partial \ln \mathcal{L}(\hat{\beta}, \hat{\alpha}_o, 0)}{\partial \hat{\alpha}_o} = -\frac{1}{2}\sum_{i=1}^{n} \frac{h'(\hat{\alpha}_o)}{h(\hat{\alpha}_o)} + \frac{1}{2}\sum_{i=1}^{n} \frac{h'(\hat{\alpha}_o)(Y_i - X_i'\hat{\beta})^2}{h(\hat{\alpha}_o)^2} = 0$$

$$\begin{aligned}\frac{\partial \ln \mathcal{L}(\hat{\beta}, \hat{\alpha}_o, 0)}{\partial \alpha_1} &= -\frac{1}{2}\sum_{i=1}^{n} \frac{h'(\hat{\alpha}_o)}{h(\hat{\alpha}_o)} Z_{1i} + \frac{1}{2}\sum_{i=1}^{n} \frac{h'(\hat{\alpha}_o)(Y_i - X_i'\hat{\beta})^2}{h(\hat{\alpha}_o)^2} Z_{1i} \\ &= \frac{1}{2}\frac{h'(\hat{\alpha}_o)}{h(\hat{\alpha}_o)} \sum_{i=1}^{n} \left[\frac{e_i^2}{h(\hat{\alpha}_o)} - 1\right] Z_{1i}\end{aligned}$$

$$\frac{\partial \ln \mathcal{L}(\hat{\beta}, \hat{\alpha}_o, 0)}{\partial \alpha_2} = \frac{1}{2}\frac{h'(\hat{\alpha}_o)}{h(\hat{\alpha}_o)} \sum_{i=1}^{n} \left[\frac{e_i^2}{h(\hat{\alpha}_o)} - 1\right] Z_{2i}$$

Breusch-Pagan LM Test

- The idea of LM test is to look at the nonzero element of above FOC to see how nonzero it is.
- More specifically, LM test is essentially based on the extent to which $e_i^2/h(\hat{\alpha}_o) - 1$ are correlated with the Z_i.
- A general computational approach can be done by running the following regression,

$$\frac{e_i^2}{\hat{\sigma}^2} = \delta_0 + \delta_1 Z_{1i} + ... + \delta_p Z_{pi} + \nu_i, \tag{1}$$

where $\hat{\sigma}^2 = e'e/n$.

Breusch-Pagan LM Test

- The BP-LM statistic can be constructed as:

$$LM = \frac{ESS}{2} \xrightarrow{d} \mathcal{X}^2(p),$$

where ESS is obtained from (1).

- Equivalently,

$$LM = \frac{g'Z\left(Z'Z\right)^{-1}Z'g - n}{2} \xrightarrow{d} \mathcal{X}^2(p),$$

where g is a vector of $e_i^2/\hat{\sigma}^2$ and Z is the $n \times (p+1)$ matrix.

Breusch-Pagan LM Test (*usual procedure*)

- Note the FOCs could be written as:

$$\frac{\partial \ln \mathcal{L}(\hat{\beta}, \hat{\alpha}_o, 0)}{\partial \alpha_{-o}} = \tilde{c} Z' \tilde{f}$$

- Under the null, $\tilde{c} Z' \tilde{f}$ is a consistent estimator of $cZ'f$.
- We can verify that

$$\frac{1}{\sqrt{n}} cZ'f \xrightarrow{d} \mathcal{N}(0, \text{plim}[\text{var}(\frac{1}{\sqrt{n}} cZ'f)])$$

- One can construct that

$$(cZ'f)'[\text{var}[cZ'f]]^{-1}(cZ'f) \xrightarrow{d} \mathcal{X}^2(p)$$

- Replacing the unknowns by consistent estimates leads to

$$LM = (\tilde{c} Z' \tilde{f})'[\text{var}[\tilde{c} Z' \tilde{f}]]^{-1}(\tilde{c} Z' \tilde{f}) \xrightarrow{d} \mathcal{X}^2(p)$$

Koenker Test

- The BP test designed under normality is actually sensitive to the normality assumption.
- It uses the fact that under the null the variance of ε_i^2 is $2\sigma^4$.
- Koenker (1981) proposed an alternative way of testing the null hypothesis that is more robust is to use nR^2 from the regression,

$$e_i^2 = \delta_0 + \delta_1 Z_{1i} + \ldots + \delta_p Z_{pi} + \varsigma_i.$$

- Koenker showed that his test statistic $(= nR^2)$ will be a Chi-squared distribution with p degrees of freedom.
- Under the normality assumption the BP test will be equivalent to Koenker's test. This is confirmed by the following Theorem.

Koenker Test

Theorem

Under the normality, one can show $\frac{ESS}{2} \to nR^2$.

Proof.

Note that

$$\begin{aligned}
\frac{ESS}{2} &= \frac{1}{2}R^2[\sum_{i=1}^{n}(\frac{e_i^2}{\hat{\sigma}^2} - 1)^2] \\
&= \frac{1}{2}R^2[\frac{1}{\hat{\sigma}^4}\sum_{i=1}^{n} e_i^4 - \frac{2}{\hat{\sigma}^2}\sum_{i=1}^{n} e_i^2 + n] \\
&\simeq \frac{1}{2}R^2[3n - 2n + n] = nR^2.
\end{aligned}$$

□

White Test

- White (1980) suggested an information matrix test which is based on comparing the two estimates,

$$\hat{\sigma}^2 \left(X'X\right)^{-1} \quad \text{and} \quad \left(X'X\right)^{-1} X'SX \left(X'X\right)^{-1},$$

  where $S = \text{diag}(e_i^2)$, to see what extent the OLS and White estimates differ.
- White's test is quite similar to Koenker's test and White's test statistic is also asymptotically $\mathcal{X}^2\left(p\right)$ under the null hypothesis of no heteroskedasticity.

White Test

- The procedure of White test could be simplified to regress the OLS residuals on original regressors along with second powers and cross product. Then calculate nR^2 from that regression to be the test statistic.
- If the Z_i vector in Koenker's test consists of total regressors used in White test, White test will be identical to Koenker's test.

White Test

Example

If the regressors are $(1, X_{1i}, X_{2i}, X_{3i})$, the White test is to run the following regression,

$$\begin{aligned} e_i^2 = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i} + \delta_4 X_{1i}^2 + \delta_5 X_{2i}^2 + \delta_6 X_{3i}^2 \\ + \delta_7 X_{1i} X_{2i} + \delta_8 X_{2i} X_{3i} + \delta_9 X_{1i} X_{3i} + \xi_i. \end{aligned} \tag{2}$$

Compute the nR^2 from regression (2) to test heteroskedasticity. What's the degrees of freedom?

Example

What if X_{3i} is a dummy variable?

Three Major Ways

- We have three major ways to correcting heteroskedasticity.
- Eicker-White correction
- GLS with known variance matrix
- FGLS with unknown variance matrix

Eicker-White Correction

- The var-cov matrix of OLS estimator under heteroskedasticity is,

$$\text{var}[\hat{\beta}] = \sigma^2 \left(X'X\right)^{-1} X'\Omega X \left(X'X\right)^{-1}.$$

- $\text{var}[\hat{\beta}]$ is not feasible due to the term $X'\Omega X$.
- One may try to estimate

$$\sigma^2 X'\Omega X = \sum_{i=1}^{n} X_i X_i' \sigma_i^2 = \sum_{i=1}^{n} X_i X_i' \text{E}[\varepsilon_i^2 | X_i]$$

- Eicker (1963) and White (1980) suggest to replace $\sigma^2 X'\Omega X$ by $X'\text{diag}(e_i^2)X \equiv X'SX$.
- We can form the popular Eicker-White covariance matrix estimator by

$$\text{var}[\hat{\beta}^W] = \left(X'X\right)^{-1} X'SX \left(X'X\right)^{-1}. \tag{3}$$

Eicker-White Correction

- White (1980) shows that (3) or Heteroskedasticity Consistent Covariance Matrix Estimaor (HCCME) is truly consistent in the sense that

$$\text{plim}\left[\left(\frac{X'X}{n}\right)^{-1}\frac{X'SX}{n}\left(\frac{X'X}{n}\right)^{-1}\right]=\sigma^2Q^{-1}\text{plim}\left(\frac{X'\Omega X}{n}\right)Q^{-1}.$$

- With HCCME in hand, it implies that all t tests and Wald tests are asymptotically valid if one uses White standard errors.
- HCCME: White correction or White washing.
- One could show that under homoskedasticity,

$$\mathrm{E}[e_i^2]=\sigma^2(1-X_i'(X'X)^{-1}X_i)=\sigma^2(1-h_{ii}).$$

  - i.e., OLS squared error is downward biased.

Eicker-White Correction – Finite-sample Refinement

- Finite-sample or small-sample bias? size distortion?
- A number of researchers have devoted to the improvement w.r.t. White correction.
- Hinkley (1977): degree of freedom correction

$$\text{var}[\hat{\beta}^H] = \left(\frac{n}{n-k}\right)\left(X'X\right)^{-1} X'\text{diag}(e_i^2)X\left(X'X\right)^{-1}.$$

- Horn, Horn, and Duncan (1975): "almost unbiased" estimator for $\sigma^2(x_i)$

$$\text{var}[\hat{\beta}^{HHD}] = \left(X'X\right)^{-1} X'\text{diag}\left(\frac{e_i^2}{1-h_{ii}}\right)X\left(X'X\right)^{-1}.$$

Eicker-White Correction – Finite-sample Refinement

- MacKinnon and White (1985): Jackknife method

$$\text{var}[\hat{\beta}^{MW}] = \frac{n-1}{n}\left(X'X\right)^{-1} X'(\text{diag}(e_i^{*2}) - \frac{1}{n}e^*e^{*\prime})X\left(X'X\right)^{-1},$$

where $e_i^* = e_i/(1-h_{ii})$ and e^* is a $n \times 1$ column vector of e_i^*.

  - Monte Carlo result of MW?
  - Stata can do this: `regress wage edu, vce(hc2)`

- Cribari-Neto (2003): removing high leverage points

$$\text{var}[\hat{\beta}^{CN}] = \left(X'X\right)^{-1} X'\text{diag}\left(\frac{e_i^2}{(1-h_{ii})^{\delta_i}}\right) X\left(X'X\right)^{-1},$$

where $\delta_i = \min\{4, nh_{ii}/\sum_{i=1}^n h_{ii}\}$.

Influential Observation

- Recall that OLS residuals are not independent and do not have the same variance (*c.f.* error term)
- More precisely, if $\text{var}[\varepsilon_i|X_i] = \sigma^2$, then one has $\text{var}[e_i] = \sigma^2(1 - h_{ii})$
- Since $\text{var}[e_i] = \sigma^2(1 - h_{ii})$, observations with large h_{ii} will have small value of $\text{var}[e_i]$, and hence tend to have residuals e_i close to zero
- h_{ii} measures the leverage of observation i

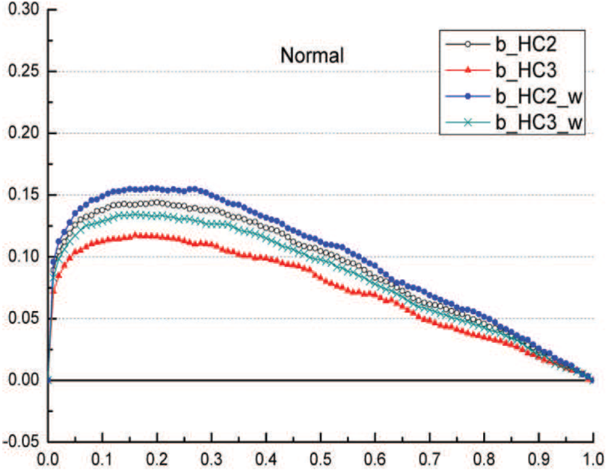
Influential Observation

- The trace of $P(\equiv X(X'X)^{-1}X')$ matrix is k (# of regressors)
- High leverage h_{ii} typically means two or three times larger than average hat value k/n
- Leverage only depends on X but not on Y
- Good/bad leverage points: high leverage points with typical/unusual Y_i
- Davidson and MacKinnon (2004): a large h_i could be influential:

$$\hat{\beta}_{-i}^{OLS} = \hat{\beta}^{OLS} - (X'X)^{-1}X_i \frac{e_i}{1-h_{ii}}, \tag{4}$$

where $\hat{\beta}^{OLS}$ and $\hat{\beta}_{-i}^{OLS}$ denote the OLS estimator using full sample, and OLS estimator using sample excluding the i^{th} observation, respectively.

Size Distortion (Error Rejection Probability)



GLS with Known Variance Matrix

- Suppose the variance function is given by

$$\text{var}\left[\varepsilon_i | Z_i\right] = \sigma^2 g\left(Z_i\right),$$

  where $g\left(\cdot\right)$ is a known function of variables Z_i that may include functions of the regressors.
- Recall the group mean regression. We know the sample size for each group, i.e., n_c is known. The variance is $\text{var}[\bar{\varepsilon}_c] = \sigma^2 / n_c$.
- To apply the GLS method, need to find the ω matrix, which satisfies $\text{var}[\omega\varepsilon] = \sigma^2 I_n$.
- The ω matrix to be applied in GLS method is trivial

$$\omega = \begin{bmatrix} g\left(Z_1\right) & 0 & \dots & 0 \\ 0 & g\left(Z_2\right) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & g\left(Z_n\right) \end{bmatrix}^{-1/2} = \text{diag}(g\left(Z_i\right)^{-1/2}).$$

GLS with Known Variance Matrix

- GLS or Aitken estimator is given by,

$$\hat{\beta}_{GLS} = \left(X^{*\prime}X^{*}\right)^{-1} X^{*\prime}Y^{*}.$$

- WLS or summation form is,

$$\hat{\beta}_{WLS} = \left(\sum_{i=1}^{n} \frac{X_i X_i'}{g\left(Z_i\right)}\right)^{-1} \left(\sum_{i=1}^{n} \frac{X_i Y_i}{g\left(Z_i\right)}\right) = \hat{\beta}_{GLS}.$$

- Recall the generalized or weighted RSS is

$$\min_{\beta} \sum_{i=1}^{n} \frac{(Y_i - X_i'\beta)^2}{g\left(Z_i\right)} = (Y - X\beta)'[\sigma^2 \mathrm{diag}(g\left(Z_i\right))]^{-1}(Y - X\beta).$$

GLS with Unknown Variance Matrix

- In most cases, the skedastic function is unknown.
- The general variance structure is:

$$\mathrm{var}[\varepsilon_i|Z_i] = g\left(Z_i, \alpha\right)$$

- Three popular settings are
  1. $g\left(Z_i, \alpha\right) = Z_i'\alpha$
  2. $g\left(Z_i, \alpha\right) = \exp\left(Z_i'\alpha\right)$
  3. $g\left(Z_i, \alpha\right) = \sigma^2\left(Z_i'\alpha\right)^2$
- We can do a two-step procedure to obtain Feasible GLS estimator.
- Estimate skedastic function first step and perform GLS procedure the second step.

GLS with Unknown Variance Matrix

- The first step is to do OLS and obtain OLS residuals e_i. Then run the regression, (assume linear one for simplicity here)

$$e_i^2 = \alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_p Z_{pi} + \xi_i$$

and obtain the fitted value to estimate $g\left(Z_i, \alpha\right)$.

$$g\left(Z_i, \hat{\alpha}\right) = \hat{\alpha}_0 + \hat{\alpha}_1 Z_{1i} + ... + \hat{\alpha}_p Z_{pi}.$$

- Note that trimming may be needed. (why?)
- The second step is to perform GLS procedure using the fitted variance terms,

$$\begin{aligned}
\hat{\beta}_{FGLS} &= \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1}Y \\
&= \left(\sum_{i=1}^{n} \frac{X_i X_i'}{g\left(Z_i, \hat{\alpha}\right)}\right)^{-1} \sum_{i=1}^{n} \frac{X_i Y_i}{g\left(Z_i, \hat{\alpha}\right)}.
\end{aligned}$$

GLS with Unknown Variance Matrix

- Need to assume the skedastic function in first step.
  - Not robust to misspecification
- Modern econometrics almost always sticks on OLS estimator and fixes up the standard errors
- However, we have seen the inefficiency of OLS estimator. Any other way to improving efficiency?
- Lin (2005) proposes a nonparametric method (series or kernel) to estimate skedastic function.
  - He also suggests an approximate MSE criterion to pick smoothing parameters.
- Using the nonparametric estimate of skedastic function to form FGLS estimator – semi-parametric approach
- Lin and Chou (2012, 2015) extend the HCCME-type finite-sample refinement to non-linear (GMM) models

Modified HCCME for Nonlinear GMM Models

Lin and Chou (2018): "Finite-sample refinement of GMM approach to nonlinear models under heteroskedasticity of unknown form," *Econometric Reviews*

Definition (Quasi-Hat Matrix)

To construct a bias-corrected estimator of the error variance for nonlinear models, we first develop a procedure for bias-reduction under homoskedasticity, and $\mathbf{W}_n$ is thus set by $(\mathbf{Z}'\mathbf{Z}/n)^{-1}$. Accordingly, for the resulting GMM estimator $\hat{\beta}_n^{ini}$, using (**??**), we have

$$(\hat{\beta}_n^{ini} - \beta_o) = \left(\mathbf{M}_o' Q_Z^{-1} \mathbf{M}_o\right)^{-1} \mathbf{M}_o' Q_Z^{-1} \frac{\mathbf{Z}'\varepsilon(\beta_o)}{n} + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{5}$$

where $Q_Z :=$plim $\mathbf{Z}'\mathbf{Z}/n$. For the sake of notational simplicity, we further define $e_i := \varepsilon_i(\hat{\beta}_n^{ini})$ and $\varepsilon_i := \varepsilon_i(\beta_o)$, respectively. Thus, using Equation (**??**), the squared residual can be represented as:

$$\begin{aligned} e_i^2 = \varepsilon_i^2 &+ \nabla_\beta g(x_i;\bar{\beta}_n)'(\hat{\beta}_n^{ini} - \beta_o)(\hat{\beta}_n^{ini} - \beta_o)' \nabla_\beta\, g(x_i;\bar{\beta}_n) \\ &- 2 \nabla_\beta\, g(x_i;\bar{\beta}_n)'(\hat{\beta}_n^{ini} - \beta_o)\varepsilon_i. \end{aligned} \tag{6}$$

Replacing $\bar{\beta}_n$ with β_o for approximation & adopting the 1st-order Taylor expansion...

Bayesian Interpretations of HCCME-type Refinements

Lin and Chou (2012): "A Note on Bayesian Interpretations of HCCME-type Refinements for Nonlinear GMM Models," *Economics Letters*

Theorem

If we use the quasi-hat matrix defined and set the prior parameter $\underline{v}_i$ *by* $\underline{v}_{HCs,i} = (1-\hat{H}_i)^{-\eta_{s,i}} - 1$ *such that the corresponding posterior parameter* $\bar{v}_{HCs,i} = (1-\hat{H}_i)^{-\eta_{s,i}}$, *it is easy to show that the approximation of* $E[\beta(\theta)|D_n]$ *is:*

$$\beta^*(\bar{\theta}_{HCs}) = \beta^*\left(\frac{\mathcal{H}_s}{\sum_{i=1}^n (1-\hat{H}_i)^{-\eta_{s,i}}}\right),$$

where $\mathcal{H}_s = ((1-\hat{H}_1)^{-\eta_{s,1}}, \ldots, (1-\hat{H}_n)^{-\eta_{s,n}})'$, *and the approximation of* $\mathrm{var}[\beta(\theta)|D_n]$ *is:*

$$V_{HCs} = \mathcal{J}_s \mathcal{C}(\mathcal{H}_s)' \mathbf{Z}' \mathrm{diag}\left(\frac{u_1^*(\bar{\theta}_{HCs})^2}{(1-\hat{H}_1)^{\eta_{s,1}}}, \ldots, \frac{u_n^*(\bar{\theta}_{HCs})^2}{(1-\hat{H}_n)^{\eta_{s,n}}}\right) \mathbf{Z}\mathcal{C}(\mathcal{H}_s).$$

where $\mathcal{J}_s = \left(1+\sum_{i=1}^n (1-\hat{H}_i)^{-\eta_{s,i}}\right)^{-1} \sum_{i=1}^n (1-\hat{H}_i)^{-\eta_{s,i}}$. *In addition,* $\eta_{2,i} = 1$, $\eta_{3,i} = 2$, $\eta_{4,i} = \delta_i$, *and* $\eta_{5,i} = \alpha_i$, *where* δ_i *and* α_i *are defined in Lin and Chou (2012).*

Serial Correlation

- Autocorrelation or serial correlation occurs when errors for different observations are correlated.
- It is most often associated with time series data because there is a natural order in the data by time period.
- The errors in time series model are always persistent, i.e., a large shock may follow if there is a large positive shock or the reverse.
- Autocorrelation could also occurs because one has omitted an important variable and the variable itself is correlated across time.
  - i.e., consider the individual specific effects in panel data

Stationarity (1)

- A univariate time series model is one that focuses on the variable ε_t alone and essentially specifies the way that ε_t depends on its prior values. Sometimes ε_t is called an innovation.
- A stochastic process is said to be covariance stationary (or weakly stationary) if the covariance only depends on the number of periods separating the elements and not where the elements are so that,

$$\text{cov}\left(\varepsilon_t, \varepsilon_{t-s}\right) = \text{cov}(\varepsilon_t, \varepsilon_{t+s}) = \gamma_s,$$

  where the auto-covariance does not depend on t.
- We can also observe that $\gamma_o = \sigma^2 = \text{var}[\varepsilon_t]$.

Stationarity (2)

- A much stronger concept of stationarity
- A stochastic process is strongly stationary if for any subset $(t_1, t_2, ..., t_k)$ of N and any real number h such that $t_i + h \in N^*$,

$$F\left(\varepsilon_{t_1}, \varepsilon_{t_2}, ..., \varepsilon_{t_k}\right) = F\left(\varepsilon_{t_1+h}, \varepsilon_{t_2+h}, ..., \varepsilon_{t_k+h}\right),$$

  where $F(\cdot)$ is the joint distribution function of the k values.
- Joint distribution function of the k values in a process ε_t is the same regardless of the origin, t_1, in the time scale.

Ergodicity

- A stationary process is ergodic if for any two bounded functions $f : R^k \rightarrow R$ and $g : R^l \rightarrow R$,

$$\lim_{n \rightarrow \infty} \left| \mathrm{E}\left[f\left(\varepsilon_t, \ldots, \varepsilon_{t+k}\right) g\left(\varepsilon_{t+n}, \ldots, \varepsilon_{t+n+l}\right)\right]\right|$$
$$= \left| \mathrm{E}\left[f\left(\varepsilon_t, \ldots, \varepsilon_{t+k}\right)\right]\right| \left| \mathrm{E}\left[g\left(\varepsilon_{t+n}, \ldots, \varepsilon_{t+n+l}\right)\right]\right| .$$

- In practice, ergodicity is usually assumed theoretically, and it is impossible to test it empirically.

Large Sample Theory With Dependent Data

- With stationarity and ergodicity we could consider the estimation of regression coefficients in time series framework.
- That is, the appropriate LLN (Ergodic LLN) and CLT (under so called mixing conditions) will be applied in deriving consistency and asymptotic normality.
    - Ergodic Theorem: If Z_t is a process that is stationary and ergodic and $\mathsf{E}[|Z_t|] < \infty$, and if $\bar{Z}_T = T^{-1}\sum_{t=1}^{T} Z_t$, then

$$\bar{Z}_T \xrightarrow{as} \mu,$$

where $\mu = \mathsf{E}[Z_t]$.
- Will discuss those tools in more advanced course.

Autocorrelation Function

- Autocorrelation function under covariance stationarity is defined by,

$$\operatorname{corr}(\varepsilon_t, \varepsilon_{t-s}) = \frac{\gamma_s}{\gamma_o}.$$

- Many possibilities exist and particular patterns in variance matrix are associated with particular special time series processes for ε_t.

White Noise

- To describe two well-known processes in time series models, it is convenient to introduce the fundamental (unobserved error) variable u_t (White noise) and to assume for simplicity that,

$$u_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_u^2).$$

- Note that normality is not really needed.

AR(p) Process

- A popular process is an Auto-Regressive of order p or AR(p) process of the form,

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \ldots + \rho_p \varepsilon_{t-p} + u_t,$$

with the special case being the AR(1) process,

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t.$$

- The auto-covariance function has the properties that,

$$\gamma_o = \sigma_u^2/(1-\rho^2) = \mathrm{var}[\varepsilon_t]$$
$$\gamma_s = \rho^{|s|}\sigma_u^2/(1-\rho^2)$$

- What's the variance covariance matrix of AR(1) process?

MA(q) Process

- Another popular process is an Moving-Average of order q or MA(q) process of the form,

$$\varepsilon_t = u_t + \lambda_1 u_{t-1} + ... + \lambda_q u_{t-q},$$

with the special case being the MA(1) process,

$$\varepsilon_t = u_t + \lambda u_{t-1}.$$

- The auto-covariance function has the properties that,

$$\begin{aligned} \gamma_o &= \sigma_u^2(1+\lambda^2) \\ \gamma_1 &= \lambda\sigma_u^2 \\ \gamma_s &= 0 \text{ for } s > 1. \end{aligned}$$

- What's the variance covariance matrix of MA(1) process?

Non-Stationarity

- Consider the following regression:

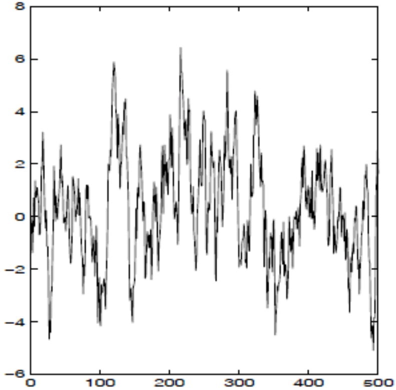
$$Y_t = \beta Y_{t-1} + \varepsilon_t$$

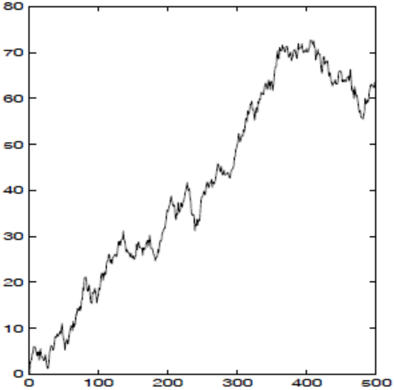
$$T(\hat{\beta} - 1) \Rightarrow \frac{1}{2} \frac{[w(1)^2 - 1]}{\int_0^1 w(r)^2 \mathrm{d}r}$$

- Last two decades, macroeconometricians developed a unique set of tools for dealing with the non-stationary processes.
- The tools are based on so called Brownian motion.
- They also found that several economic variables are non-stationary.
  - Unit root, co-integration, and spurious regression issues
- C. W. Granger and R. Engle won the 2003 Nobel Prize due to their contribution on time series area.

Non-Stationarity vs. Stationary

- $Y_t = 0.9 Y_{t-1} + \varepsilon_t$ vs. $Y_t = Y_{t-1} + \varepsilon_t$





A Motivating Example

- As we know OLS is unbiased and consistent even errors are non-spherical.
- It implicitly implies that there are no lagged dependent variables.
- Consider the model with lagged dependent variable and AR(1) errors.

$$\begin{aligned} Y_t &= \beta Y_{t-1} + \varepsilon_t \\ \varepsilon_t &= \rho \varepsilon_{t-1} + u_t. \end{aligned}$$

- One can show that OLS estimator $\hat{\beta} = \sum Y_t Y_{t-1} / \sum Y_{t-1}^2$ is inconsistent.

$$\begin{aligned} \text{plim}\hat{\beta} &= \beta + \frac{\text{plim} \sum Y_{t-1}\varepsilon_t / T}{\text{plim} \sum Y_{t-1}^2 / T} = \beta + \frac{\text{cov}\left[Y_{t-1}, \varepsilon_t\right]}{\text{var}\left[Y_{t-1}\right]} \\ &= \beta + \frac{\rho\left(1-\beta^2\right)}{1+\beta\rho}. \end{aligned}$$

A Motivating Example

- To show the inconsistency of OLS estimator, note that:

$$\begin{aligned} \mathrm{cov}\left[Y_{t-1}, \varepsilon_t\right] &= \mathrm{cov}\left[Y_{t-1}, \rho\varepsilon_{t-1} + u_t\right] = \rho\mathrm{cov}\left[Y_{t-1}, \varepsilon_{t-1}\right] \\ \mathrm{cov}\left[Y_t, \varepsilon_t\right] &= \mathrm{cov}\left[\beta Y_{t-1} + \varepsilon_t, \varepsilon_t\right] = \beta\mathrm{cov}\left[Y_{t-1}, \varepsilon_t\right] + \sigma_\varepsilon^2 \\ \mathrm{var}\left[Y_t\right] &= \beta^2\mathrm{var}\left[Y_{t-1}\right] + 2\beta\mathrm{cov}\left[Y_{t-1}, \varepsilon_t\right] + \sigma_\varepsilon^2 \end{aligned}$$

- So that,

$$\begin{aligned} \mathrm{plim}\hat{\beta} &= \beta + \frac{\mathrm{plim}\sum Y_{t-1}\varepsilon_t / T}{\mathrm{plim}\sum Y_{t-1}^2 / T} = \beta + \frac{\mathrm{cov}\left[Y_{t-1}, \varepsilon_t\right]}{\mathrm{var}\left[Y_{t-1}\right]} \\ &= \beta + \frac{\rho\left(1 - \beta^2\right)}{1 + \beta\rho}. \end{aligned}$$

How to Test for Autocorrelation

- Consider the regression model without lagged dependent variables is,

$$Y_t = X_t'\beta + \varepsilon_t$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t.$$

- The most well-known test for AR(1) errors is a test of,

$$H_o : \rho = 0$$
$$H_1 : \rho \neq 0.$$

- One can use the Durbin-Watson test.

Durbin-Watson Test

- The rationale behind the Durbin-Watson (DW) test is that under H_o,

$$\frac{\mathrm{E}[\varepsilon_t - \varepsilon_{t-1}]^2}{\mathrm{E}\left[\varepsilon_t^2\right]} = \frac{\mathrm{E}[\varepsilon_t^2 + \varepsilon_{t-1}^2 - 2\varepsilon_t\varepsilon_{t-1}]}{\mathrm{E}\left[\varepsilon_t^2\right]} = \frac{2\sigma_\varepsilon^2 - 2\rho\sigma_\varepsilon^2}{\sigma_\varepsilon^2} = 2\,(1-\rho) = 2.$$

- When the alternative is true DW will deviates from 2 and can be either close to 0 or 4 depending on the sign of ρ.
- The DW test is based on the sample analog of the parameter computed using the OLS residuals,

$$d = \frac{\sum_{t=2}^{T} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{T} e_t^2} \simeq 2\,(1-r)\,,$$

where r is the regression coefficient in the following,

$$e_t = \rho e_{t-1} + error.$$

DW Critical Values

- The big problem of DW test statistic is that the exact null distribution of d depends on the data matrix X and therefore cannot be tabulated in general.
  - that is, $d = 2\,(1 - r) \overset{d}{\nrightarrow} \mathcal{N}(.)$
- However, it has been shown that the null distribution of d lies between the distribution of a lower bound $(d - lb)$ and upper bound $(d - ub)$.
- The good thing is that the distributions of $(d - lb)$ and $(d - ub)$ are independent of X. Thus, the critical values for $(d - lb)$ and $(d - ub)$ could be tabulated.

DW under Positive Alternatives

- Let d_l and d_u denote the critical values of $(d - lb)$ and $(d - ub)$.
- Under the test,

$$H_o : \rho = 0$$
$$H_1 : \rho > 0,$$

the decision rule is,

1. Do not reject the null if $d > d_u$.
2. Reject the null if $d \leq d_l$.
3. Inconclusive test if $d_l < d \leq d_u$.

DW under Negative Alternatives

- Under the test,

$$H_o : \rho = 0$$
$$H_1 : \rho < 0,$$

the decision rule is,

1. Do not reject the null if $d < 4 - d_u$.
2. Reject the null if $d \geq 4 - d_l$.
3. Inconclusive test if $4 - d_u < d \leq 4 - d_l$.

Difficulties of DW Test

- Even we could compare the DW statistic with d_l and d_u, which are usually tabulated at the end of statistic textbook. We still have several difficulties.
  1. We may yield inconclusive conclusion.
  2. DW test applies only when X contains a constant term and for non-stochastic X.
  3. DW test could not handle the model with lagged dependent variables.
  4. DW test is designed for test of AR(1).
- The first difficulty could be resolved by some econometric packages such as Shazam, which provides the exact DW distribution and exact p-values.
- The second difficulty could be referred to Farebrother (1980).
- The last two difficulties could be resolved later on.

Durbin's h Test

- If the regression model contains lagged dependent variables, DW statistic has been shown to be biased.

$$Y_t = \gamma Y_{t-1} + X_t'\beta + \varepsilon_t$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t.$$

- One must use an adapted version known as Durbin's h statistic,

$$h = (1 - \frac{d}{2})\sqrt{\frac{T}{1 - T[s.e.(\hat{\gamma})^2]}} \sim \mathcal{N}(0,1),$$

  where $s.e.(\hat{\gamma})$ is the standard error of the (inconsistent?) OLS estimator of the coefficient on lagged dependent variable.
- It is easy to implement because the Durbin's h is distributed as $\mathcal{N}(0,1)$. No inconclusive area!
- The obvious drawback is that Durbin's h could not be calculated if $s.e.(\hat{\gamma})^2 \geq 1/T$.

Wallis Test

- If one is interested in testing the periodic (quarterly) effect that is causing the errors across the same periods but in different periods to be correlated, the Wallis test for AR (4) could be employed.
- The AR(4) specification is:

$$\varepsilon_t = \rho_4 \varepsilon_{t-4} + u_t$$

- And we would like to test

$$H_o : \rho_4 = 0$$

- Wallis statistic is given by,

$$d_4 = \frac{\sum_{t=5}^{T} \left(e_t - e_{t-4}\right)^2}{\sum_{t=1}^{T} e_t^2},$$

where the critical values are presented in Wallis (1972) or in [JD] Appendix D-6.

Box-Pierce Q and Liung-Box Q Tests

- The Box-Pierce Q statistic based on the first p autocorrelation coefficients is given by,

$$Q = T \sum_{j=1}^{p} r_j^2 \xrightarrow{d} \mathcal{X}^2(p) \text{ under the null,}$$

where

$$r_j = \frac{\sum_{t=j+1}^{T} e_t e_{t-j}}{\sum_{t=1}^{T} e_t^2}.$$

- Liung-Box Q prime statistic is:

$$Q' = T(T+2) \sum_{j=1}^{p} r_j^2 / (T-j)$$

  - a revised Box-Pierce Q statistic
  - better small-sample performance

Breusch-Godfrey LM Test for AR(p) or MA(q)

- The hypothesis is,

$$H_o : \text{no autocorrelation}$$
$$H_1 : \varepsilon_t = \text{AR}(p) \text{ or } \varepsilon_t = \text{MA}(q).$$

- Note that under H_1, our model could be written as,

$$Y_t = X_t'\beta + \rho_1\varepsilon_{t-1} + ... + \rho_p\varepsilon_{t-p} + u_t \quad \text{or}$$
$$Y_t = X_t'\beta + \theta_1 u_{t-1} + ... + \theta_p u_{t-p} + u_t.$$

- Serial correlation influences the conditional mean by adding extra variables which we estimate with $e_{t-1}, e_{t-2}, ..., e_{t-p}$ under the null.

Breusch-Godfrey LM Test for AR(p) or MA(q)

- The test for above hypothesis is much like a Breusch-Pagan test for heteroskedasticity. The same test is used for either structure. Here are the steps.
  1. (Obtain OLS residual, e_t) Run the auxiliary regression,
  $$e_t = X_t'\alpha + \theta_1 e_{t-1} + ... + \theta_p e_{t-p} + error, \tag{7}$$
  and let BG $= TR^2$, where R^2 is the usual R-square from (7).
  2. Reject the null if BG is larger than a critical values from the $\mathcal{X}^2\,(p)$.
- [JD] page 185 uses the AR(1) case to illustrate Breusch-Godfrey test.
  - setup the log-likelihood: $\ln \mathcal{L}$
  - compute the score and information matrix
  - evaluate at the restricted estimates (X_t is not restricted)
  - LM type statistics is equivalent to compute TR^2 from (7).

Possible Remedies

- HACCME
- FGLS
  - Prais-Winsten Procedure
  - Iterated Prais-Winsten Procedure
  - Cochrane-Orcutt Procedure
  - Iterated Cochrane-Orcutt Procedure
  - Hildreth-Lu Procedure
- MLE

Correcting Standard Errors

- The variance covariance matrix of OLS is,

$$\text{var}[\hat{\beta}] = \sigma^2 \left(X'X\right)^{-1} X'\Omega X \left(X'X\right)^{-1}.$$

- The difficulty is that Ω is usually unknown. One may try to estimate

$$\begin{aligned}\sigma^2 X'\Omega X &= \sum_{t=1}^{T}\sum_{s=1}^{T} X_t X_s' \mathrm{E}\left[\varepsilon_t \varepsilon_s\right] \\ &= \sigma^2 \sum_{t=1}^{T}\sum_{s=1}^{T} X_t X_s' \rho_{|t-s|}.\end{aligned}$$

- How to estimate the so-called "long run variance"?
- Adjusting Standard Errors by Newey-West (1987)

Newey and West Approach

- The approach suggested by Newey and West (1987) is similar to White approach in that, one uses

$$\widehat{NW}_T = \sum_{t=1}^{T} X_t X_t' e_t^2 + \sum_{j=1}^{L} \sum_{t=j+1}^{T} \omega(j)\, e_t e_{t-j}(X_t X_{t-j}' + X_{t-j} X_t'),$$

  where $\omega(j) = 1 - j/(L+1)$ are kernel weights that ensure that the variance covariance matrix is positive semi-definite and the value L is the lag truncation.
- How many autocorrelations one should use? It is generally hard to pick. One may look at autocorrelation function of the residuals for help. Andrews (1991) offers a method for the selection of L.
- How to pick the kernel function?
  - Bartlett kernel
  - Quadratic spectrum kernel
  - Triangular kernel

HACCME (HAC)

- The resulting estimator known as Newey-West Heteroskedasticity and Autocorrelation Consistent Co-variance Matrix Estimator (HACCME) or simply HAC takes the form,

$$\text{var}[\hat{\beta}^{NW}] = (X'X)^{-1}\,\widehat{NW}_T\,(X'X)^{-1}.$$

- Newey and West showed that provided $L \to \infty$,

$$\text{plim}\left(\frac{X'X}{n}\right)^{-1}\left(\frac{\widehat{NW}_T}{n}\right)\left(\frac{X'X}{n}\right)^{-1} = \sigma^2 Q^{-1} Q_{\Omega} Q^{-1}.$$

- This implies that all t tests and Wald tests are asymptotically valid if one uses the Newey-West standard errors. Note that Newey-West estimator is robust for both Heteroskedasticity and Autocorrelation.

Covariance Structure

- Assume that we have an AR(1) error structure:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

- The variance covariance structure of AR(1) is given by:

$$\sigma^2\Omega = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & 1 & \dots & \rho^{T-2} \\ \dots & \dots & \dots & \dots \\ \rho^{T-1} & \rho^{T-2} & \dots & 1 \end{bmatrix}.$$

Prais-Winsten Method

- To implement GLS procedure, we need to find the ω matrix.

$$\omega = \frac{1}{\sigma_{\varepsilon}^{2}} \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} & \dots & 0 & 0 \\ \dots & \dots & \dots & \frac{1}{\sqrt{1-\rho^2}} & \dots \\ 0 & 0 & \dots & -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{bmatrix}.$$

Prais-Winsten Method

- One can do Cholesky decomposition with respect to a symmetric positive definite matrix.
- Decompose Ω into LU with lower and upper triangular matrices.
- Note that $U' = L$.
- Once we have L and U. Ω^{-1} will be straightforward.
- Commands for doing this. `chol`(X), `cholesky`(X)
- Note that any matrix that is a constant proportion to ω could also serve as a transformation matrix.
- Consider

$$\dot{\omega} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 & 0 \\ -\rho & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & 1 & \dots \\ 0 & 0 & \dots & -\rho & 1 \end{bmatrix}.$$

Prais-Winsten Method

- Consider
$$X = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$
- We have `chol(X)`:
$$\begin{bmatrix} \sqrt{2} & 0 & 0 \\ -\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2}\sqrt{3} & 0 \\ 0 & -\frac{1}{3}\sqrt{2}\sqrt{3} & \frac{2}{3}\sqrt{3} \end{bmatrix} \begin{bmatrix} \sqrt{2} & -\frac{1}{2}\sqrt{2} & 0 \\ 0 & \frac{1}{2}\sqrt{2}\sqrt{3} & -\frac{1}{3}\sqrt{2}\sqrt{3} \\ 0 & 0 & \frac{2}{3}\sqrt{3} \end{bmatrix}$$

Prais-Winsten Method

- The transformed model takes the form:

$$\dot{\omega} Y = \dot{\omega} X \beta + \dot{\omega} \varepsilon,$$

which implies,

$$\begin{aligned} Y_1^* &= \sqrt{1-\rho^2} Y_1 = \sqrt{1-\rho^2} X_1' \beta + \sqrt{1-\rho^2} \varepsilon_1 \\ Y_t^* &= Y_t - \rho Y_{t-1} = (X_t - \rho X_{t-1})' \beta + \varepsilon_t - \rho \varepsilon_{t-1}. \end{aligned}$$

- It is clear that we now have spherical errors:

$$\begin{aligned} \operatorname{var}\left[Y_1^*\right] &= (1-\rho^2)\sigma_\varepsilon^2 \\ \operatorname{var}[Y_t^*] &= (1-\rho^2)\sigma_\varepsilon^2 \\ \operatorname{cov}\left[Y_t^*, Y_{t-i}^*\right] &= 0 \end{aligned}$$

Prais-Winsten Method

- Similar problem to heteroskedasticity!
- Need to estimate ρ. Recall the D-W statistic. Can do FGLS!
- Prais-Winsten procedure or Iterated Prais-Winsten procedure.
  1. Run the OLS and obtain OLS residual e_t.
  2. Run the regression,
     $$e_t = \rho e_{t-1} + error$$
     and obtain the estimate $\hat{\rho}$.
  3. Run the following regression to obtain the estimate $\hat{\beta}_{FGLS}$.
     $$\begin{aligned} \sqrt{1-\hat{\rho}^2}Y_1 &= \sqrt{1-\hat{\rho}^2}X_1'\beta + \sqrt{1-\hat{\rho}^2}\varepsilon_1 \\ Y_t - \hat{\rho}Y_{t-1} &= \left(X_t - \hat{\rho}X_{t-1}\right)'\beta + \varepsilon_t - \hat{\rho}\varepsilon_{t-1} \end{aligned} \tag{8}$$
  4. One may obtain the residuals,
     $$\hat{e}_t = Y_t - X_t'\hat{\beta}_{FGLS}$$
     and repeat the steps 2 and 3 until convergence has been achieved.

Cochrane-Orcutt Method

- If one drop the first observation in (8), the procedure is called Cochrane-Orcutt (1949) or Iterated Cochrane-Orcutt estimator.
- In large samples, PW and CO procedures are likely to be the same.
- In finite samples, since CO procedure omits the first observation, it is usually less satisfactory.

Hildreth-Lu Procedure

- The Hildreth-Lu procedure is basically doing the grid search to find the $\rho \in (-1, 1)$ that minimizes the RSS of the model.

$$\hat{\beta}_{FGLS}(\rho) = \arg\min \left(1 - \rho^2\right) \left(Y_1 - X_1'\beta\right)^2 + \sum_{t=2}^{T} (Y_t - \rho Y_{t-1} - (X_t - \rho X_{t-1})' \beta)^2 \tag{9}$$

- However, Beach and MacKinnon (1978) argue that H-L procedure is very inefficient due to the computational burden.

Maximum Likelihood Approach

- To go through the maximum likelihood estimation, we use the fact that,

$$f(Y_1, Y_2, ..., Y_T) = f(Y_1)\, f(Y_2|Y_1) \ldots f(Y_T|Y_1, Y_2, ..., Y_{T-1})\,.$$

- We have,

$$\begin{aligned} \sqrt{1-\rho^2}\,Y_1 &\sim \mathcal{N}(\sqrt{1-\rho^2}X_1'\beta, \sigma_u^2) \\ Y_t &\sim \mathcal{N}(\rho Y_{t-1} + (X_t - \rho X_{t-1})'\,\beta, \sigma_u^2), \end{aligned}$$

so that,

$$f(Y_1, Y_2, ..., Y_T) = f(Y_1)\, f(Y_2|Y_1) \ldots f(Y_T|Y_{T-1})\,.$$

Maximum Likelihood Approach

- Therefore,

$$f(Y_1) = \sqrt{1-\rho^2}\frac{1}{\sqrt{2\pi\sigma_u^2}}\exp[-\frac{1-\rho^2}{2\sigma_u^2}(Y_1 - X_1'\beta)^2]$$

$$f(Y_t|Y_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_u^2}}\exp[-\frac{(Y_t - \rho Y_{t-1} - (X_t - \rho X_{t-1})'\beta)^2}{2\sigma_u^2}].$$

- The log-likelihood function is,

$$\begin{aligned}\ln \mathcal{L}(\beta,\rho,\sigma_u^2) = & -\frac{T}{2}\left[\ln(2\pi) + \ln(\sigma_u^2)\right] + \frac{1}{2}\ln(1-\rho^2) \\ & - \frac{\left(1-\rho^2\right)\left(Y_1 - X_1'\beta\right)^2}{2\sigma_u^2} \\ & - \frac{\sum_{t=2}^{T}(Y_t - \rho Y_{t-1} - (X_t - \rho X_{t-1})'\beta)^2}{2\sigma_u^2}.\end{aligned} \tag{10}$$

MLE vs. FGLS

- Note the MLE is just GLS estimator provided ρ and σ_u^2 are known. If ρ and σ_u^2 are unknown, log-likelihood (10) will be nonlinear in nature and should be solved by nonlinear optimization routines.
- Recall that we regress e_t on e_{t-1} to obtain the estimate of ρ based on OLS estimator of β.
- It is equivalent to minimize the term in (10),

$$\sum_{t=2}^{T}(Y_t - \rho Y_{t-1} - (X_t - \rho X_{t-1})' \beta)^2,$$

  where the other terms involving ρ have been ignored and thus $\hat{\rho}$ is not MLE.
- This means that the iterated procedure (FGLS procedure) will not produce MLE.